Investigating possible causes of bias in a progress test translation

Cecilio-Fernandes, Dario; Bremers, André; Collares, Carlos Fernando; Nieuwland, Wybe; Vleuten, Cees van der; Tio, René A

Published in:
Korean journal of medical education

DOI:
10.3946/kjme.2019.130

Publication date:
2019

ORIGINAL RESEARCH

Investigating possible causes of bias in a progress test translation: an one-edged sword

Dario Cecilio-Fernandes[1,2], André Bremers[3], Carlos Fernando Collares[4], Wybe Nieuwland[5], Cees van der Vleuten[4] and René A. Tio[4,6]

[1]School of Medical Sciences, University of Campinas, Campinas, Brazil, [2]Center for Education Development and Research in Health Professions (CEDAR), University Medical Center Groningen, University of Groningen, Groningen, [3]Department of Surgery, Radboud University Nijmegen Medical Center, Nijmegen, [4]Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences at Maastricht University, Maastricht, [5]Department of Cardiology, University Medical Center Groningen, University of Groningen, Groningen, and [6]Department of Cardiology, Catharina Hospital, Eindhoven, The Netherlands

Purpose: Assessment in different languages should measure the same construct. However, item characteristics, such as item flaws and content, may favor one test-taker group over another. This is known as item bias. Although some studies have focused on item bias, little is known about item bias and its association with items characteristics. Therefore, this study investigated the association between item characteristics and bias.

Methods: The University of Groningen offers both an international and a national bachelor's program in medicine. Students in both programs take the same progress test, but the international progress test is literally translated into English from the Dutch version. Differential item functioning was calculated to analyze item bias in four subsequent progress tests. Items were also classified by their categories, number of alternatives, item flaw, item length, and whether it was a case-based question.

Results: The proportion of items with bias ranged from 34% to 36% for the various tests. The number of items and the size of their bias was very similar in both programmes. We have identified that the more complex items with more alternatives favored the national students, whereas shorter items and fewer alternatives favored the international students.

Conclusion: Although nearly 35% of all items contain bias, the distribution and the size of the bias were similar for both groups. The findings of this paper may be used to improve the writing process of the items, by avoiding some characteristics that may benefit one group whilst being a disadvantage for others.

Key Words: Educational measurement, Bias, Medical education

Introduction

Progress testing is a longitudinal assessment of students' knowledge development by periodical testing at end level [1]. The progress test has been used as a benchmark tool, for comparison either within the same university [2] or between universities, both nationally [3-5] and internationally [6,7]. Although the progress test is a reliable and valid tool to measure students' knowledge growth [1,3,4], one precondition for its application is that it will only detect differences in the

Received: July 10, 2019 • Revised: July 17, 2019 • Accepted: July 17, 2019
Corresponding Author: Dario Cecilio-Fernandes (https://orcid.org/0000-0002-8746-1680)
Center for Education Development and Research in Health Professions (CEDAR), University of Groningen and University Medical Center Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands
Tel: +31.50.363.8415 email: d.cecilio.fernandes@umcg.nl

Korean J Med Educ 2019 Sep; 31(3): 193-204.
https://doi.org/10.3946/kjme.2019.130
eISSN: 2005-7288

level of knowledge and skilled knowledge application of students. Any other influence leading to bias should be avoided. Or at least, such influences should be acknowledged and quantified to ascertain fair judgement of students and to render reliable program information regarding the education quality in medical schools.

Item bias will, by definition, favor a subgroup or be detrimental to another subgroup [8]. Such bias may result either in failure of students to pass the tests due to other reasons than lack of knowledge or vice versa. Identifying items that are biased is usually done after the test was taken by means of psychometric analysis of the items. Consequently, items that are biased may be deleted. Excluding items however may impact negatively on the coverage of content of the test and hence its validity [9]. Although it is important to identify items with bias for the sake of quality control and fair judgement, thorough understanding of possible sources of bias may benefit test validity. From the literature, it is known that there may be several sources of item bias, like language, category of the items, and item flaws.

Tests that are available in different languages should be measuring the same construct to allow a meaningful comparison [10-12]. A poor translation can compromise the validity of the test, making it difficult to compare both scores because the two test forms may not be construct equivalent [10]. Consequently, effectively reducing the language barriers in assessment would reduce the loss of the content validity, resulting in a fair assessment for all students [13-15]. Research in non-native English speakers has shown that students' performance on a knowledge test in English may be worse: due to insufficient proficiency in English: the test becomes a language test [16]. Students' knowledge cannot be assessed adequately if students do not understand the vocabulary and linguistic structures [15]. Thus, the test score becomes a variable, dependent on knowledge and on English proficiency. The content validity of the test is at stake. Although research has shown that language is a specific source of item bias [17,18], hereafter called item language bias, it may be unjustified to blame this item bias on language factors alone. Different languages are often associated with cultural differences which, by themselves, may constitute an additional source of bias [19,20].

Traditionally, studies considering item bias focused mostly on verifying whether an item presents a bias by comparing two or more subgroups. However, item bias may also be influenced by other factors such as item content and item flaws, which in turn may also be related to language. Zenisky et al. [9] for example found that items that were related to earth and space science, physical sciences, and technology presented bias favoring males compared to females. In the medical literature, Swanson et al. [21] identified that longer items would benefit female students, though the effect size was small. They also revealed that bias may be related to the item categories, which were classified as internal medicine, obstetrics- gynecology, pediatrics, psychiatric, and surgery. Whilst obstetric-gynecological, pediatric, and psychiatric items favored female students, surgical items favored the males.

The relation between item flaws and item bias has not been studied sufficiently yet, but item flaws may certainly compromise the validity of a test [9]. Items that contain writing flaws were shown to be up to 15% more difficult than items that are perfect in this respect [22]. Moreover, the flawed items are more likely to penalize skilled students than borderline students [23]. The effect is sufficiently significant to influence the pass failing decision [22,23]. Though adding to construct-irrelevance variance, these flawed items had little effect on the psychometric properties of the test [22] and therefore may remain unidentified.

Language, item content, and item flaws can have a significant impact on the validity and fairness of assessment. However, very little is known about the association between items with bias regarding language (national versus international students) and other sources of bias, since most studies have focused merely on identifying items with bias. Therefore, we raised the following research questions: (1) Which items exhibit bias when comparing national and international students? (2) Is there an association between items' characteristics (i.e., item content) and item language bias?

To analyze whether an item was biased, we used the differential item function (DIF) analysis. DIF analysis tests whether test-takers of two or more subgroups would have the same probability of answering an item correctly when they have the same level of ability [8]. More precisely, if an item parameter differs across groups, an item displays DIF. DIF is a robust method that considers difference at every ability level [8,11,24]. DIF analysis has been extensively used to investigate item bias when comparing male versus female, native versus non-native speakers and white population versus minority, including in the context of medical education (for example, see Hope et al. [24]). Since DIF analysis has a basis in the item response theory, we also investigated the assumptions of unidimensionality and local independency, as a requirement of the item response theory. The item response theory is a mathematical model that establishes a relation between the knowledge or ability of the test taker, the difficulty of the test items and the probability of a correct answer. The item response theory based method estimates student ability (θ) and item difficulty [25,26].

Methods

1. Setting

Since 2009, the University Medical Center Groningen has been offering a national and an international bachelor's degree program in medicine. Both programs' teaching methods are based on the problem-based learning curriculum, sharing the same learning goals, content, and material. The international track is taught in English, whereas the regular program is taught in the Dutch language which is the native language to the vast majority of students.

Although the admission requirements are the same for both groups, all international students take a proficiency in English test (IELTS, International English Language Testing System), except the native English speakers who are the minute minority of students. Adequate scientific level is assured by 1 year of pre-university education if candidates fail an entrance test in this subject. Both programs are regulated by the same rules and cutoff scores, assuring a comparable level of students' knowledge in both tracks.

We used data from the University of Groningen concerning students' scores on four Dutch Interuniversity Progress Tests of Medicine, including students from the first 3 years of medical training (bachelor) [1,27]. Students in the regular track (hereafter called national track) answered the questions in Dutch and students in the other track (hereafter called international track) answered the same questions in English.

2. Progress test

The Dutch Interuniversity Progress Test of Medicine contains 200 multiple-choice questions based on the Dutch National Blueprint for the Medical Curriculum

[28] and it is administered 4 times a year. The Dutch Progress Test uses formula scoring: if students do not know the answer, they can choose a "question mark" option. To correct for guessing, the incorrect answers receive a penalty to outbalance the chance of scoring by guessing. The penalty varies according to the number of alternatives, which ranges from 2 to 5 alternatives. The penalty therefore ranges from -1.00 to -0.25, respectively (-1/[number of alternatives-1]). A correct answer is rewarded with 1 point and the "question mark" scores zero.

The Dutch Progress Test is administered in two languages, Dutch and English. The Dutch Progress Test is translated into English by a native speaker who is an official certified translator with years of experience in translating medical documents and tests. Subsequently, the English translation is revised by a physician how is a native English-speaker. The chairman of the Dutch Progress Test consortium overseas the translation process and review final version of the English progress test. There is no back translation.

3. Data analysis

Before describing the analysis of DIF and the sources of bias, we will describe the calibration and preliminary analyses.

1) Calibration

We analyzed 800 questions in four subsequent progress tests from both programs. We analyzed the data using the Rasch Partial Credit Model for polytomous categories because the categories follow an ordinal arrangement. The right answer has the highest value (6); the question mark having the second highest value (5); and the penalties having the lowest values, representing the amount of penalty based on the formula scoring (4, 3, 2, and 1) [26].

2) Preliminary analysis

Unidimensionality was tested with Principal-Components Analysis of Residuals from the Rasch Model and a fit only approach [29]. For the Principal-Components Analysis of Residuals, another dimension would be considered when having more than two items. If another dimension had more than two items, we compared the amount of explained variance of both dimensions. If another dimension is presented, the progress test could be measuring another construct than medical knowledge. For the "fit only" approach, the two fit parameters, infit and outfit, for the item and person were assessed to test unidimensionality. For both parameters, the optimal fit value is 1.00 [30] with a range from 0.50 to 1.50 [31]. If the parameter for the items exceeds 2.0, this is considered to be a threat to the validity of the test [31] and the item should preferably be excluded from the test.

Local independency was estimated by the correlation of the standardized residual, which analyzes how much of the variance is common of two items. When two items share more than half of their variance, they may be measuring similar content. Therefore, only one of the two items is needed for the test. Local independency can be assumed adequate when items present a residual correlation lower than 0.7 [32]. Local independency assures that there is no pattern in the residuals, meaning that parts of the data that were not explained by the model are not related. Also, when local independence is violated, it may inflate the estimation of the item difficulty. Finally, an overlap between too many item pairs, with high correlation of Rasch residuals, may be due to the occurrence of multidimensionality.

3) Differential item function

We used three criteria to determine whether the item has DIF: (1) a value higher than 2.4 in the t-test [33], (2) a significant probability of t, and (3) a significant difference calculated by Mantel-Haenszel method [34]. We considered an item to display DIF when an item met each of all three criteria. Subsequently, we assess the

size of the DIF as suggested by Zwick et al. [35]. When the difference between the DIF of both groups is smaller than 0.43, is considered negligible; from 0.43 to 0.64 it is considered slight to moderate, and higher than 0.64 is considered moderate to large. Negligible degree of DIF is often disregarded since it does not affect the score.

4) Sources of DIF

Five variables were investigated as possible sources of DIF. (1) Category of the items: the items in progress test are divided in 17 categories: respiratory system, blood & immune system, musculoskeletal system, mental health care, reproductive system, pregnancy, childbirth & puerperium, cardiovascular system, hormones & metabolism, endocrine system, dermis & connective tissue, personal and social aspects, digestive/gastrointestinal system, nutritional disorders nervous system & senses, kidneys & urinary system, molecular & cellular aspects, epistemology, methodology & applied biostatistics, stages of life, knowledge of skills, and preventive medicine. (2) Number of alternatives: the alternative options ranged from 2 to 5 per question. (3) Item flaw: items were classified as flawed when one or more of the following problems were presented: logical clues, greater details in the correct answer, implausible distractors, unfocused items, no correct or more than one correct answer, unnecessary information, unbalanced distractors, and negative items. The items were classified by one of the co-authors (R.T.) who is an experienced reviewer of the items of the progress test and former chairman of the Dutch Progress Test Committee. The items flaws were extracted for the literature and the categorization followed the guidelines of writing items for medical education [22]. (4) Item length: the words of each questions were counted. (5) Case-based questions: questions were classified as simple questions when there was no patient, and vignette questions when a patient was presented.

The items were calibrated, and DIF was calculated using Winsteps ver. 3.70.1.1 (Winsteps Co., Beaverton, USA). Descriptive analyses and inferential statistics were calculated using the IBM SPSS ver. 25.0 (IBM Corp., Armonk, USA).

4. Ethical statement

Ethical approval was not sought since reanalysis of historical data is automatically ruled exempt. This exemption is because our data that was collected as part of an existing educational assessment (progress test) without the necessity of collecting new data. All data were anonymized and handled with confidentiality. We also conduct our work following the Declaration of Helsinki and the privacy policy of the University of Groningen.

Results

We gathered progress test data from 5,186 bachelor students. From those, 907 were students who attended the international track and 4,279 who attended the national track.

1. Preliminary analysis

The first residual contrast (dimension) after obtaining the Rasch measures, had more than two items for all tests, indicating that a second dimension may have been present. The variance explained by the items was more than 7 times the variance explained of the first contrast: 22.6% versus 3.0%. Moreover, the variance explained in the first contrast was smaller than the variance explained by persons and items, meaning that the amount of the variance explained by the "extra" dimension is negligible. These findings indicate that the four progress tests may be unidimensional, since comparable values were found for the four tests.

The fit parameters of the item were in the optimal interval, i.e., between 0.50 and 1.50 [31] and the mean values were near 1.00, which is the optimal value for the infit and outfit. There was only one item in test 1 that had outfit higher than 2.00. The values of mean, standard deviation, minimum and maximum of measurement, infit, outfit, and error, based on Rasch outcomes, can be visualized in Table 1.

For the person parameters, there were some violations of the maximum and minimum value of the recommended interval. However, those violations are acceptable, since measuring students' ability may also be related to other factors, such familiarly with the test, cheating or items being answered using a methodological approach or answered exceptionally slowly.

Considering both the Principal-Components Analysis of Residuals and the only fit approach, all four progress tests can be considered unidimensional, meeting the first assumption of the Rasch Model.

The highest correlation of the standardized residual was 0.54; thus, the local independency holds, since items present a correlation lower than 0.7. This indicates that the second assumption of the Rasch Model is met, indicating that items are locally independent. Since the two assumptions were met, Rasch Model is suitable for the data analysis.

Table 1. Mean, SD, Minimum and Maximum of Measurement, Infit, Outfit, and Error for Items and Person

Test	Category	Items				Person			
		Measure	Infit	Outfit	Error	Measure	Infit	Outfit	Error
Test 1 (September)	Mean±SD	0.00±1.39	1.00±0.13	0.92±0.30	0.09±0.03	-1.93±1.14	0.99±0.13	0.92±0.38	0.23±0.07
	Minimum	-3.78	0.73	0.40	0.06	-5.54	0.68	0.18	0.72
	Maximum	2.93	1.58	2.06	0.24	0.78	1.79	5.83	0.17
Test 2 (December)	Mean±SD	0.00±1.37	1.00±0.12	0.96±0.26	0.08±0.04	-1.41±0.93	0.99±0.12	0.96±0.28	0.20±0.04
	Minimum	-3.42	0.73	0.49	0.06	-4.27	0.69	0.36	0.17
	Maximum	3.87	1.53	1.75	0.32	0.83	1.54	4.21	0.43
Test 3 (February)	Mean±SD	0.00±1.28	1.00±0.14	0.97±0.28	0.08±0.03	-1.31±0.91	0.99±0.10	0.97±0.27	0.19±0.03
	Minimum	-3.99	0.71	0.48	0.06	-3.72	0.73	0.34	0.16
	Maximum	3.73	1.61	1.90	0.29	1.28	1.42	3.13	0.36
Test 4 (May)	Mean±SD	0.00±1.28	1.00±0.11	1.00±0.22	0.08±0.03	-1.10±0.86	0.99±0.11	1.00±0.25	0.19±0.03
	Minimum	-3.38	0.74	0.62	0.06	-3.74	0.70	0.39	0.16
	Maximum	3.52	1.32	1.85	0.25	0.92	1.65	3.31	0.36

SD: Standard deviation.

Table 2. Number of Items That Presented Differential Item Function Favoring the National or International Track Divided by the Following Categories: Negligible, Moderate, and Larger

Test	Category	Size			
		Negligible (%)	Moderate (%)	Larger (%)	Total (%)
Test 1 (September)	International	4 (2)	7 (3.5)	25 (12.5)	36 (18)
	National	16 (8)	8 (4)	8 (4)	32 (16)
Test 2 (December)	International	9 (4.5)	4 (2)	25 (12.5)	38 (19)
	National	21 (10,5)	3 (1.5)	9 (4.5)	33 (16.5)
Test 3 (February)	International	4 (2)	8 (4)	24 (12)	36 (18)
	National	17 (8.5)	3 (1.5)	11 (5.5)	31 (15.5)
Test 4 (May)	International	8 (4)	4 (2)	24 (12)	36 (18)
	National	17 (8.5)	4 (2)	9 (4.5)	30 (15)
Total	International	25 (3.12)	23 (2.87)	98 (12.25)	146 (18.5)
	National	71 (8.8)	18 (2.25)	37 (4.62)	126 (15.75)

2. Differential item function

Items that presented differential item functioning ranged from 66 (34%) to 71 (36%) items of the 200 in each test. Although items were favoring both groups, 146 items (54% of the items with DIF) favored the international students and 126 items (46% of the items with DIF) favored the national students. This indicates that international students have a higher probability of answering a question correctly than national students with the same level of knowledge. Most of the items (72.6%) with larger size DIF were favoring the international students, whereas most of items (74%) with negligible DIF were favoring the national students. The items with moderate DIF seems to have a similar distribution between national and international track (see details on Table 2). More importantly, the distribution shows that there was no systematic bias against any group, since the bias occurred for groups concurrently, indicating that the final score was unlikely to be affected by the bias.

3. Sources of differential item function

1) Category of the items

The distribution of questions with DIF was similar in nine of the 17 categories (Table 3). From the other eight categories, four favored the international track: (1) cardiovascular system; (2) hormones & metabolism and endocrine system; (3) digestive/gastrointestinal system, nutritional disorders; and (4) molecular & cellular aspects. The categories that favored the national track were: (1) mental health care; (2) personal and social

Table 3. Distribution of Items in the 17 Categories of the Progress Test

Category	Items						Total of items
	No DIF		DIF favoring: international		DIF favoring: national		
	No. of items (%)	Min/max size	No. of items (%)	Min/max size	No. of items (%)	Min/max size	
Respiratory system	46 (74.4)	0.01/1.9	12 (19.7)	0.13/1.21	3 (4.9)	0.34/1.04	61
Blood & immune system	27 (67.5)	0/2.4	4 (10)	0.02/1.34	9 (22.5)	0.02/0.97	40
Musculoskeletal system	30 (61.2)	0.03/2.36	11 (22.4)	0.22/1.59	8 (16.3)	0.21/0.85	49
Mental health care	23 (50)	0.02/1.95	7 (15.2)	0.13/1.54	16 (34.8)	0.03/1.25	46
Reproductive system, pregnancy, childbirth & puerperium	25 (58.1)	0.19/2.56	11 (25.6)	0.28/1.3	7 (16.3)	0.08/0.23	43
Cardiovascular system	39 (65)	0.02/2.3	16 (26.7)	0.54/2.08	5 (8.3)	0.13/0.89	60
Hormones & metabolism, endocrine system	25 (64.1)	0.03/2.24	13 (33.3)	0.38/2.4	1 (2.6)	1.33/1.33	39
Dermis & connective tissue	23 (60.5)	0.11/2.18	8 (21.1)	0.45/1.81	7 (18.4)	0.10/1.01	38
Personal and social aspects	21 (44.7)	0.01/2.31	6 (12.8)	0.07/1.02	20 (42.6)	0.11/1.85	47
Digestive/gastrointestinal system, nutritional disorders	33 (68.8)	0.01/2.56	12 (25)	0.24/1.06	3 (6.3)	0.01/1.08	48
Nervous system & senses	46 (83.6)	0.05/2.56	6 (10.9)	0.29/1.12	3 (5.5)	0.13/0.7	55
Kidneys & urinary system	52 (68.4)	0.01/2.56	14 (18.4)	0.17/1.61	10 (13.2)	0.02/0.69	76
Molecular & cellular aspects	24 (68.6)	0.03/1.87	8 (22.9)	0.03/1.5	3 (8.6)	0.36/0.62	35
Epistemology, methodology & applied biostatistics	23 (65.7)	0.17/1.97	3 (8.6)	0.85/1.09	9 (25.7)	0.21/1.04	35
Stages of life	18 (66.7)	0.07/1.29	4 (14.8)	0.43/1.22	5 (18.5)	0.06/1.35	27
Knowledge of skills	44 (74.6)	0/2.5	8 (13.6)	0.62/1.95	7 (11.9)	0.04/1.92	59
Preventive medicine	11 (45.8)	0.05/0.96	3 (12.5)	0.08/2.22	10 (41.7)	0.02/0.86	24

DIF: Differential item function, Min: Minimum, Max: Maximum.

aspects; (3) epistemology, methodology & applied bio-statistics; and (4) preventive medicine.

2) Number of alternatives

It seems that the number of alternatives has an impact on items with DIF. While items with two alternatives (n=54) favored the international track more (37%) than the national track (7.4%), items with five alternatives (n=29) favored the national track (20.7% versus instead 6.9%). Items with three (n=234) and four (n=465) alternatives seem to have similar impact on items with DIF for both tracks.

3) Item flaws

In total, although 147 (18.8%) items presented a writing flaw, most items presented implausible distractors (n=50) and unbalanced distractors (n=25). Of the 147 items, only 41 (5.24%) presented DIF. From those 41, 21 favored the international track and 20 the national track. Looking in more detail at the different categories of flawed items, the distribution between items favoring national and international track was similar (Table 4).

4) Item length

We found that the items that favored the international track (M=22.46) were significantly shorter than the items that favored national track (M=28.45, t=-2.734; $p<0.05$).

5) Case-based questions

Questions were classified as non-case-based questions (n=526) and case-based questions (n=256). Although the distribution of question with DIF was similar to both

Table 4. Number of Items Favoring the National or International Track Divided by Item Flaws

Variable	Category	No DIF		DIF favoring: international		DIF favoring: national		Total
		No. of items (%)	Min/max size	No. of items (%)	Min/max size	No. of items (%)	Min/max size	
Logical clues	No	507 (65.3)	0.00/2.56	145 (18.7)	0.02/2.4	125 (16.1)	0.01/1.92	777
	Yes	3 (60)	0.23/1.47	1 (20)	0.80/0.80	1 (20)	1.06/1.06	5
Greater detail in correct option	No	504 (65.3)	0.00/2.56	143 (18.5)	0.02/2.4	125 (16.2)	0.01/1.92	772
	Yes	6 (60)	0.04/2.36	3 (30)	0.91/2.08	1 (10)	0.69/0.69	10
Implausible distractors	No	475 (64.9)	0.00/2.56	140 (19.1)	0.02/2.40	117 (16)	0.01/1.92	732
	Yes	35 (70)	0.01/1.87	6 (12)	0.55/0.92	9 (18)	0.13/1.25	50
Unfocused stem	No	500 (65)	0.00/2.56	145 (18.9)	0.02/2.40	124 (16.1)	0.01/1.92	769
	Yes	10 (76.9)	0.20/1.95	1 (7.7)	1.08/1.08	2 (15.4)	0.25/1.1	13
No correct or more than one correct answer	No	504 (65.8)	0.00/2.56	142 (18.5)	0.02/2.40	120 (15.7)	0.01/1.92	766
	Yes	6 (37.5)	0.01/0.80	4 (25)	0.56/1.22	6 (37.5)	0.03/1.06	16
Unnecessary information	No	502 (65.1)	0.00/2.56	145 (18.8)	0.02/2.40	124 (16.1)	0.01/1.92	771
	Yes	8 (72.7)	0.11/1.95	1 (9.1)	1.08/1.08	2 (18.2)	0.06/0.69	11
Unbalance in distractors	No	493 (65.1)	0.00/2.56	141 (18.6)	0.02/2.40	123 (16.3)	0.01/1.92	757
	Yes	17 (68)	0.01/1.91	5 (20)	0.47/1.21	3 (12)	0.02/0.41	25
Negative items	No	496 (64.8)	0.00/2.56	144 (18.8)	0.02/2.40	125 (16.3)	0.01/1.92	765
	Yes	14 (82.4)	0.02/1.31	2 (11.8)	0.35/0.49	1 (5.9)	0.31/0.31	17

DIF: Differential item function, Min: Minimum, Max: Maximum.

Table 5. Number of Items Favoring the National or International Track Divided by the Number of Non- and Case-Based Questions

Case-based questions	No DIF		DIF favoring: international		DIF favoring: national		Total
	No. of items (%)	Min/max size	No. of items (%)	Min/max size	No. of items (%)	Min/max size	
No	314 (59.7)	0/2.56	120 (22.8)	0.2/2.4	92 (17.5)	0.01/1.85	526
Yes	196 (76.6)	0/2.56	26 (10.2)	0.07/2.08	34 (13.3)	0.02/1.92	256

DIF: Differential item function, Min: Minimum, Max: Maximum.

tracks (Table 5), non-case-based questions (40.3%) seem to be more likely to present DIF than case-based questions (23.4%).

Discussion

In this study, we sought to identify biased items and investigate whether there was a pattern in the item characteristics that may have caused the bias. Although there was a high percentage of biased items, those biased items favored the national and the international students in the same proportion. Contrary to our findings, the literature shows that biased items usually favor one subgroup more than another [11,12,15-17].

We found that the long items seem to favor the national track. Although the educational literature focuses more on comparing native and non-native speakers sitting in a test in the same language, we believed that a parallel can be drawn. Usually long items favored the native speakers [36,37] since the complexity of the language is often higher in longer items compared to short ones. This complies with our results since the native speakers in our study were almost exclusively present in the national track. The assumption of longer items favoring the national track is also supported by the number of alternatives: the questions with five alternatives have favored the national group, but questions with two alternatives have favored the international group. Furthermore, two of the categories (mental health care and personal and social aspects) that favored the national track are considered as the most complex in terms of language when compared to the other categories. Longer items and more alternatives may also be indicative of more subtilty in the questions, which may explain why they are more difficult for non-native speakers. Thus, it seems that item's characteristics may also be a source of bias, especially when considering the linguistic complexity and their length.

Identifying language bias and its association with the source of that bias, is crucial for quality control. More practically, the information regarding the sources may help improve the process of test development. For example, one may choose to have tests with only three and four options, since using items with three and four options may equally distribute the numbers of items with bias across various groups of students. Also, one may consider writing shorter items, which, in turn, may be hard when writing case-based questions. However, it seems that non-case-based questions have higher percentages of items with DIF than case-based questions. Interestingly, item flaws have little impact on the number of items with DIF and the distribution was similar across both groups. Although other studies have suggested that flawed items may have impact on students' scores [20,21], we found no evidence for an advantage, neither for the national nor the international track. Yildirim and Berberoğlu [38] suggested that reviewing the items considering the possible sources of DIF would decrease the number of items that presented DIF. Thus, taking the findings of this study into account when reviewing items, may help to decrease bias.

This study has a few limitations. Although the difference between samples in international and national tracks is large, Winsteps gives different weights to different samples, allowing correction for sample size differences. Furthermore, the international sample was enough to calibrate and have stable parameters using Rasch. For a two-tailed 99% confidence intervals, the minimum sample size is 108 subjects [39]. Another limitation may be that it was not possible to differentiate whether the bias was due to the language, culture or both. Though there were a few English native speakers, the vast majority were international students for whom

English was a second language. On one hand, eliminating the English native speakers would probably increase the difference found in item bias between national and international track. On the other hand, eliminating the English native speakers implies in underrepresenting the international track and its multicultural environment. Only one researcher revised these questions, yet this researcher is an experienced reviewer of the items of the progress test and former chairman of the Dutch progress test committee. Studies focus mostly on investigating whether an item is biased, but an understanding of the cause of the bias may help us to decrease the number of items with bias.

In conclusion, in this study, we sought to understand the association between items with bias and possible source of such bias by analysing four Dutch Interuniversity Progress Tests of Medicine applied in the native language to regular students and in English to international students. Although nearly 35% of items presented bias, the distribution as well as the size of the items favoring both groups were similar. The identification of sources of bias (item category, word count and number of alternatives) may help to improve the quality control of the test development. If a test has national and international takers, the size of the items, number of alternatives should be considered. Furthermore, it seems that case-based questions may help to decrease bias, when considering the size of the questions.

ORCID:

Dario Cecilio-Fernandes: https://orcid.org/0000-0002-8746-1680;
André Bremers: https://orcid.org/0000-0002-2871-4836;
Carlos Fernando Collares: https://orcid.org/0000-0003-0914-3430;
Wybe Nieuwland: https://orcid.org/0000-0003-2829-3127;
Cees van der Vleuten: https://orcid.org/0000-0001-6802-3119;
René A. Tio: https://orcid.org/0000-0003-1164-5827

Acknowledgements: We would like to acknowledge the Dutch Working Group of the Interuniversity Progress Test of Medicine.

Funding: None.

Conflicts of interest: No potential conflict of interest relevant to this article was reported.

Author contributions: Conception and design of the study: DCF, RT; acquisition, analysis, and interpretation of the data: DCF, AB, CFC, WN, CVDV, RT; acquisition of study data: DCF, AB, CFC, WN, CVDV, RT; critical revision: AB, CFC, WN, CVDV, RT; drafting the article: DCF; and final approval of the version to be published and agreement of its publication: DCF, AB, CFC, WN, CVDV, RT.

References

1. Wrigley W, van der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE guide no. 71. Med Teach. 2012;34(9):683-697.
2. Cecilio-Fernandes D, Aalders WS, de Vries J, Tio RA. The impact of massed and spaced-out curriculum in oncology knowledge acquisition. J Cancer Educ. 2018; 33(4):922-925.
3. Muijtjens AM, Schuwirth LW, Cohen-Schotanus J, van der Vleuten CP. Differences in knowledge development exposed by multi-curricular progress test data. Adv Health Sci Educ Theory Pract. 2008;13(5):593-605.
4. De Champlain AF, Cuddy MM, Scoles PV, et al. Progress testing in clinical science education: results of a pilot project between the National Board of Medical Examiners and a US Medical School. Med Teach. 2010; 32(6):503-508.
5. Cecilio-Fernandes D, Aalders WS, Bremers AJ, Tio RA, de Vries J. The impact of curriculum design in the

acquisition of knowledge of oncology: comparison among four medical schools. J Cancer Educ. 2018;33(5): 1110-1114.
6. Albano MG, Cavallo F, Hoogenboom R, et al. An international comparison of knowledge levels of medical students: the Maastricht Progress Test. Med Educ. 1996;30(4):239-245.
7. Verhoeven BH, Snellen-Balendong HA, Hay IT, et al. The versatility of progress testing assessed in an international context: a start for benchmarking global standardization? Med Teach. 2005;27(6):514-520.
8. Lord FM. A study of item bias, using item characteristic curve theory. In: Poortinga YH, ed. Basic Problems in Cross-Cultural Psychology. Amsterdam, The Netherlands: Swets and Zeitlinger, B.V.; 1977:19-29.
9. Zenisky AL, Hambleton RK, Robin F. DIF detection and interpretation in large-scale science assessments: informing item writing practices. Educ Assess. 2004;9(1-2): 61-78.
10. Gierl MJ. Construct equivalence on translated achievement tests. Can J Educ. 2000;25(4):280-296.
11. Hulin CL. A psychometric theory of evaluations of item and scale translations: fidelity across languages. J Cross Cult Psychol. 1987;18(2):115-142.
12. Van de Vijver F, Hambleton RK. Translating tests. Eur Psychol. 1996;1(2):89-99.
13. Abedi J, Lord C, Hofstetter C. Impact of selected background variables on students' NAEP math performance. http://cresst.org/wp-content/uploads/TECH478.pdf. Published 1998. Accessed November 9, 2018.
14. Kiplinger VL, Haug CA, Abedi J. Measuring math, not reading, on a math assessment: a language accommodations study of English language learners and other special populations. Paper presented at: the Annual Meeting of the American Educational Research Association; April 24-28, 2000; New Orleans, LA, USA. https://files.eric.ed.gov/fulltext/ED441813.pdf. Accessed November 9, 2018.
15. Abedi J. Language issues in item development. In: Downing SM, Haladyna TM, eds. Handbook of Test Development. Mahwah, USA: Lawrence Erlbaum Associates Publishers; 2006:377-398.
16. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for educational and psychological testing. Washington DC, USA: American Educational Research Association; 1999.
17. Montero E. Linguistic and cultural influences on differential item functioning for Hispanic examinees in a standardized secondary level achievement test. Tallahassee, USA: The Florida State University; 1994.
18. Snetzler S, Qualls AL. Examination of differential item functioning on a standardized achievement battery with limited English proficient students. Educ Psychol Meas. 2000;60(4):564-577.
19. Byrne BM. Testing for equivalent self-concept measurement across culture: issues, caveats, and application. Int Adv Self Res. 2003;1:291-314.
20. Van de Vijver F, Tanzer NK. Bias and equivalence in cross-cultural assessment: an overview. Eur Rev Appl Psychol. 1998;47(4):263-279.
21. Swanson DB, Clauser BE, Case SM, Nungester RJ, Featherman C. Analysis of differential item functioning (DIF) using hierarchical logistic regression models. J Educ Behav Stat. 2002;27(1):53-75.
22. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract. 2005;10(2):133-143.
23. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ. 2008;42(2): 198-206.

24. Hope D, Adamson K, McManus IC, Chis L, Elder A. Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. BMC Med Educ. 2018;18(1):64.
25. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44(1):109-117.
26. Cecilio-Fernandes D, Medema H, Collares CF, Schuwirth L, Cohen-Schotanus J, Tio RA. Comparison of formula and number-right scoring in undergraduate medical training: a Rasch model analysis. BMC Med Educ. 2017;17(1):192.
27. Tio RA, Schutte B, Meiboom AA, et al. The progress test of medicine: the Dutch experience. Perspect Med Educ. 2016;5(1):51-55.
28. Van Herwaarden C, Laan R, Leunissen R. The 2009 framework for undergraduate medical education in the Netherlands. Utrecht, The Netherlands: Dutch Federation of University Medical Centres; 2009.
29. Tennant A, Pallant JF. Unidimensionality matters!: a tale of two Smiths? Rasch Meas Trans. 2006;20(1): 1048-1051.
30. Bond TG, Fox CM. Applying the Rasch Model: fundamental measurement in the human sciences. Mahwah, USA: Erlbaum; 2001.
31. Wright B, Linacre J. Reasonable mean-square fit values. Rasch Meas Trans. 1994;8(3):370.
32. Yen WM. Scaling performance assessments: strategies for managing local item dependence. J Educ Meas. 1993; 30(3):187-213.
33. Draba R. The identification and interpretation of item bias. Chicago, USA: Department of Education, Education Statistics Laboratory, The University of Chicago; 1977.
34. Holland PW, Thayer DT. Differential item performance and the Mantel-Haenszel procedure. In: Wainer H, Braun HI, eds. Test Validity. Hillsdale, USA: Lawrence Erlbaum Associates Inc.; 1988:129-145.
35. Zwick R, Thayer DT, Lewis C. An empirical Bayes approach to Mantel-Haenszel DIF analysis. J Educ Meas. 1999;36(1):1-28.
36. Sireci SG, Allalouf A. Appraising item equivalence across multiple languages and cultures. Lang Test. 2003;20(2): 148-166.
37. Allalouf A, Hambleton RK, Sireci SG. Identifying the causes of DIF in translated verbal items. J Educ Meas. 1999;36(3):185-198.
38. Yildirim HH, Berberoğlu G. Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. Int J Test. 2009;9(2):108-121.
39. Linacre J. Sample size and item calibration stability. Rasch Meas Trans. 1994;7(4):328.